THE VULNERABILITY OF THE CAPARO TEST BEFORE AI'S BLACK BOX

Raajshikha Pandit*

ABSTRACT

The Caparo test, established in Caparo Industries Plc. v. Dickman, has now been in use for over thirty years to establish the duty of care to ascertain the tort of negligence. Though pretty comprehensive, this test seems to have failed before Artificial Intelligence (AI)'s black box, which prevents humans from comprehending how AI systems make their decisions. This article provides a brief overview of the concept of duty of care in torts and a critical analysis of the applicability of the three-fold Caparo test to AI, considering the black box problem. It recognises the issues faced by courts, leading them to shove liability of AI to developers and deployers in the absence of any suitable laws. Towards the end, it also attempts to provide balanced suggestions based on a few recent legal developments to overcome the issue of ascertaining AI liability presently. This article recognises the growing importance of AI and the consequent rise in harms caused by AI, and therefore, attempts to comprehensively analyse the liability of AI under the tort of negligence.

Keywords: Caparo Test, Negligence, Artificial Intelligence, Black Box.

INTRODUCTION

It is a widely accepted fact that the human brain is a trend-finding and a prediction machine that works in rather unpredictable ways. Artificial Intelligence (AI) is directly connected with brain science. It attempts to develop systems that can perform tasks that require human intelligence; hence, it is a 'thinking machine', very similar to the human brain.[1] Associate Professor Samir Rawashdeh, specialising in AI, reveals how these deep learning systems are trained like humans, and they subsequently develop their own "neural network" to categorise

*BA LLB, FIRST YEAR, FACULTY OF LAW, UNIVERSITY OF DELHI.

[1] Jingtao Fan, Lu Fang, Jiamin Wu, Yuchen Guo, Qionghai Dai, 'From Brain Science to Artificial Intelligence' (2020) 6 (3) Engineering 248 <https://www.sciencedirect.com/science/article/pii/S2095809920300035> accessed 6 February 2026

previously unexperienced and new phenomena. In Large Language Models, neurons fire for several unrelated ideas, making the explanation of the results nearly impossible. This phenomenon is called "polysemanticity".[2] Thus, similar to human intelligence, there are often no tracks of how a deep learning system reaches certain conclusions because the initial inputs easily get lost. This inability to understand how these systems make decisions is called the "black-box problem".[3] These models are trained using vast amounts of data and complex deep learning processes, which is why the providers and deployers of AI systems, as well as those affected by their use, sometimes fail to fully understand their working. There is thus a lack of transparency in AI's outputs, which often leads to unreliable results.

The adoption of AI is growing rapidly in a variety of sectors like finance, entertainment, healthcare, etc., primarily due to its ability to process large amounts of data easily. This implies that malfunctions in AI can affect a large number of sectors in the economy and a large number of people altogether. Nowadays, almost every machine learning model, like ChatGPT, Llama, etc., is a black box AI. Such models are either created deliberately by developers and programmers to secure "intellectual property" or form into "organic black boxes" due to the complexities of the deep learning systems.[4] Models created deliberately could still be controlled and analysed, but the ones that evolve themselves become problematic when outputs of AI come under scrutiny in the courts of law. In fact, it can cause serious harm, and there are beliefs that AI will not improve our lives in the future.[5] Even robots run on software and AI, and may be connected to the Internet, but they differ from computers and software in the sense that they act upon the real world and are capable of causing physical harm to people or property.[6]

Thus, AI systems, whether in the form of prompt systems working on electronic devices or Robots, working physically, are deployed in a large number of industries in the present world. A number of key industries, in fact, have placed heavy reliance on AI systems. For example,

[2] Conor Bronsdon, 'A Review of Towards Monosemanticity: Decomposing Language Models with Dictionary Learning' (*Galileo*, 1 August 2025) <https://galileo.ai/blog/anthropic-ai-interpretability-breakthrough> accessed 16 February 2026
[3] Samir Rawashdeh, 'AI's mysterious 'black box' problem, explained' (*University of Michigan-Dearborn*, 6 March 2023) <https://umdearborn.edu/news/ais-mysterious-black-box-problem-explained> accessed 7 February 2026
[4] Matthew Kosinski, 'What is black box AI?' (*IBM Think*, 29 October 2024) <https://www.ibm.com/think/topics/black-box-ai> accessed 8 February 2026
[5] Lee Rainie and Janna Anderson, 'Code-Dependent: Pros and Cons of the Algorithm Age' (2017) Pew Research Centre <https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2017/02/PI_2017.02.08_Algorithms_FINAL.pdf > accessed 8 February 2026
[6] Ryan Calo, 'Robotics and the Lessons of Cyberlaw' (2015) 103 CALIF. L. REV. 513, 533-34

in the field of medicine, some scholars believe that machines are "consistently more accurate than human doctors", and it can be inferred that they perform better. They even disregard concerns of black box in medical AI, but claim that safety is grounded on accuracy and not explainability of AI.[7] Such views, however, get negated with every case of medical AI negligence when medical practitioners find ways to shift the blame to AI systems.

In light of these circumstances, therefore, it becomes important to delve into the present applicable laws on AI amid concerns of this black box issue, with respect to the tort of negligence.

DRAWING A RELEVANCE OF THIS PROBLEM TO THE TORT OF NEGLIGENCE

The law of torts is a Common Law that is developed through judicial precedents rather than statutes. Torts are essentially associated with establishing legal injury, duty, and the harm caused. According to Professor Winfield, a "tortious liability arises from breach of a duty primarily fixed by law". It is essentially "accident-law-plus", where the primary claims lie under negligence or strict liability. Certain decisions serving as platforms for discussions on "economic or moral theory" constitute the "plus". Thus, the law of Torts expresses how costs of accidents are allocated under Common Law, while simultaneously providing instructions on law, economics, and philosophy.[8]

Negligence is a kind of tort that works against every 'person' who breaches their duty of care towards other people. It emerged as a distinct tort around the mid-nineteenth century in Brown v. Kendall.[9]

In the earliest instances, the essence of this tort was to impose liability for carelessly causing harm to another person.[10] It was also important to establish a causality between a breach of duty and consequent damage that was "natural, probable, proximate and not too remote."[11] The concept of duty of care has been evolving over a long time, beginning with the 'neighbour

[7] Hanhui Xu, Kyle Michael James Shuttleworth, 'Medical artificial intelligence and the black box problem: a view based on the ethical principle of "do no harm"' (2024) 4 (1) Intelligent Medicine 52 <https://www.sciencedirect.com/science/article/pii/S2667102623000578 > accessed 8 February 2026
[8] John C.P. Goldberg and Benjamin C. Zipursky, Torts as Wrongs (2010) 88 Tex. L. Rev. 917, 918 <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=1672&context=faculty_scholarship> accessed February 8, 2026
[9] *Brown v. Kendall* 60 Mass. 292, 1850 WL 4572 (Mass.)
[10] David G. Owen, 'The Five Elements of Negligence' (2007) 35 (4) Hofstra Law Review 1671 <https://scholarlycommons.law.hofstra.edu/cgi/viewcontent.cgi?article=2282&context=hlr> accessed 9 February 2026
[11] Ibid

principle' in the 1932 case of Donoghue v. Stevenson,[12] with the Caparo test is the latest and most applicable case, which is why it remains the primary focus of this article. After Lord Wilberforce's two-stage test in 1978, there came Lord Keith's 'two-part test', involving "foreseeability and proximity". Thus, the problematic component of "proximity" became a part of the duty.[13] Thereafter, ascertaining justness, fairness, and reasonability also became an issue, as this third part created a lot of confusion and inconsistency.[14] The duty of care simply involves the exercise of care that a reasonable person would use under similar circumstances.[15] Professor Winfield attributes four attributes to negligence: duty, causation, breach, and damages. Various courts also recognise two or three, or even five elements, roughly covering these four plus the element of "proximate cause".[16] All these cases shall be dealt with later in this article.

As mentioned earlier, AI is tasked with replicating human capabilities, and sometimes these models go beyond human capabilities by recognising and modelling patterns that are too complex to be processed by humans.[17] In such circumstances, it becomes difficult to apply the existing tort laws on AI to establish clear liabilities. Even within torts, there is a number of them: negligence, strict liability, product liability, or insurance, that may apply to AI, and society uses the best applicable of these from an "accident-prevention perspective", primarily to achieve redressal.[18] But the current laws are not yet advanced enough to address such liability perfectly, as algorithms are not considered cognizable parties or lawful "persons" who can be sued. It is believed that it would not be prudent to grant civil rights to AI, as they are merely based on recognising patterns and feeding on data, rather than applying logic and reasoning.[19] Since no liability can be attributed to AI, it is the developers or users are the ones who are blamed. But it is important to recognise that most of the injuries caused by AI are actually due

[12] *Donoghue v. Stevenson* [1932] UKHL 100, [1932] AC 562
[13] Tan Keng Feng, 'THE THREE-PART TEST: YET ANOTHER TEST OF DUTY IN NEGLIGENCE' 1989) 31 Mal. L. R. 223, 224 <https://law.nus.edu.sg/sjls/wp-content/uploads/sites/14/2024/07/1285-1989-31-mal-dec-223.pdf> accessed 9 February 2026
[14] Pallavi Agarwal, 'Caparo Test' (2021) 2 Indian JL & Legal Rsch 1
[15] Sweety Phogat, 'Tort of Negligence in India' (2018) 6 (1) IJCRT 626 <https://ijcrt.org/papers/IJCRT1705097.pdf> accessed 10 February 2026
[16] David G. Owen, 'The Five Elements of Negligence' (2007) 35 (4) Hofstra Law Review 1671, 1672-73 <https://scholarlycommons.law.hofstra.edu/cgi/viewcontent.cgi?article=2282&context=hlr> accessed 10 February 2026
[17] Andrew D. Selbst, 'Negligence and AI's Human Users' (2020) 100 Boston University Law Review 1315, 1319 <https://www.bu.edu/bulawreview/files/2020/09/SELBST.pdf> accessed 9 February 2026
[18] Ibid at 1320
[19] Brandeis Marshall, 'No legal personhood for AI' (2023) 4 (11) Patterns <https://www.sciencedirect.com/science/article/pii/S2666389923002453#bib4> accessed 8 February 2026

to its unpredictability rather than human intervention per se.[20] These injuries are what often constitute 'harms' under the law of torts. However, there exists a clear dichotomy between the general principle of the law of torts, which is to necessarily generate a "private right of action" for its victim,[21] and the contemporary challenges to provide remedies for such "harms" caused by non-human systems. These challenges stem from the fact that traditional rules either seem to have a tendency to leave victims to tend to their harms in the absence of any laws on AI liability or shift liability to companies, even when they give vivid warnings to users regarding the use of AI. Thus, AI shall be given the status of something like an "electronic person" for the mere purpose of adjudication and establishing liability, the details of which shall be dealt with in a later section of the article.

On connecting the dots of the above discussion, it may be concluded that since AI is not considered a "person" in the eyes of the law, a victim's right to action may not be adequately catered to as per the present laws. Furthermore, AI has become an inevitable tool in professional life today. It, thus, becomes essential to provide a clear picture of the extent of the "reasonable care" required to be exercised by both the users and developers of AI, as in the absence of any such mechanism, the inevitable rise of AI systems and their occasional failures could leave a plethora of victims without any remedies for harms caused due to AI negligence. Therefore, a deeper understanding of the applicability of the tort of negligence to this "black box" problem becomes important to match the high speed of AI's growth to the slow and steady evolution of the law of torts.

RESEARCH OBJECTIVE

This article aims to break down the three-fold Caparo test to establish a duty of care under negligence and critically analyse its applicability to AI, given that it might become unforeseeable due to the Black Box problem, establishes no clear proximity between its users and developers, and also poses an issue in ascertaining the level of "reasonable care" to be exercised by a deep learning system, not a real person.

[20] Mihailis E. Diamantis, Employed Algorithms: A Labour Model of Corporate Liability for AI, 72 DUKE L.J. 797, 800-03 (2023)
[21] John C.P. Goldberg and Benjamin C. Zipursky, Torts as Wrongs (2010) 88 Tex. L. Rev. 917, 918 <https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=1672&context=faculty_scholarship> accessed February 8, 2026

RESEARCH METHODOLOGY

This article undertakes both a technical explanation of the black box problem of Artificial Intelligence, as well as an in-depth analysis of relevant laws and precedents, to provide a comprehensive view of the issue. It primarily deals with an analysis of the Caparo test to check its application to Artificial Intelligence. It has utilised research papers, blogs, newspaper articles, and case laws to provide a comprehensive analysis of the main problem. Thus, this article is based on doctrinal research, incorporating a black letter approach by first explaining the concepts in detail, then delving into their analysis, and finally providing suggestions based on the analysis of the problems.

LITERATURE REVIEW

This paper provides a comprehensive analysis of negligence with respect to AI, but fails to provide any expansion of categories of duty of care. It also does not adequately address the presence of a black box within AI.[22] This paper highlights the constraints of the black box in proving general liability of AI, and explains how explainable AI may be able to improve the determination of AI liability.[23] This paper aims to establish a standard of negligence within the current law that will help in adequately addressing AI negligence and holding companies liable in a way that is fair and progressive under the law. Despite the recognition of AI's black box, however, developers and companies are often treated as primary perpetrators.[24]

This paper provides a comprehensive analysis of the Anns test and delves into the question of policy considerations within the test. It also highlights the importance of the Caparo test, being the most suitable for negligence laws presently.[25] This paper provides a case-based analysis of the third leg of the Caparo test. It helps understand how this test might fail to address the issue of fairness, justice, and reasonableness while ascertaining AI liability.[26] This case provides an understanding of why the previous tests were insufficient to address liability in negligence. It is the most recent and most applicable case, which is why this paper seeks to check its applicability to AI. Though the Caparo test is a vital tool, it can prove to be a "human-centric"

[22] Andrew D. Selbst, 'Negligence and AI's Human Users' (2020) 100 Boston University Law Review 1315, 1319 <https://www.bu.edu/bulawreview/files/2020/09/SELBST.pdf>
[23] P.H. Padovan, C.M. Martins, Chris Reed, 'Black is the new orange: how to determine AI liability' (2022) 31 Artif Intell Law 133 <https://doi.org/10.1007/s10506-022-09308-9>
[24] Mihailis E. Diamantis, 'Reasonable AI: A Negligence Standard' (2025) 78 Vand L Rev 573
[25] Joost Blom, 'Do We Really Need the Anns Test for Duty of Care in Negligence?' (2016) 53 Alta L Rev 895
[26] Pallavi Agarwal, 'Caparo Test' (2021) 2 Indian JL & Legal Rsch 1

test being applied to "non-human" logic, which is why the applicability of this test must be redefined to suit AI needs.[27]

This paper provides an in-depth breakdown of the three-fold Caparo test and reveals its strictly human-centric approach. Thus, this also reveals how it cannot be aptly applied to AI without modifications, despite being the most comprehensive test for negligence.[28] This paper analyses the prospect of establishing vicarious liability for AI and acknowledges that there are not enough studies that examine the consequences of situations where non-transparent algorithms were themselves negligent in some circumstances. Thus, this article also seeks to formulate a suitable vicarious liability model for AI.[29] This paper provides a short analysis of the applicability of the present framework of negligence and general tortious liability on AI, without considering the black box problem.[30]

Thus, there are not enough studies that provide comprehensive analyses of the liability of AI under the tort of negligence, and almost none that examine the applicability of the presently used test to determine the duty of care under negligence with respect to AI.

CRITICAL ANALYSIS OF THE APPLICABILITY OF THE CAPARO TEST TO AI

Understanding the CAPARO Test: The duty of care was earlier restricted to being established only on a case-by-case basis, and there was no absolute standard to establish this. In Donoghue v. Stevenson, Lord Atkin formulated the "neighbour principle". He sought to reiterate the Biblical rule that one must love one's neighbour to establish that one "must take reasonable care to avoid acts or omissions" which one "can reasonably foresee would be likely to injure" one's neighbour. Here, "neighbour" does not imply people within physical proximity but those whom one can reasonably foresee to be in danger of being affected by one's negligent actions. Thus, there must be a "connection," and this connection must give rise to a duty of care. In Home Office v. Dorset Yacht Co. Ltd.,[31] Lord Diplock based 'duty of care' upon a "reasonable foreseeability of harm" and "proximity of those who were more at risk". The neighbourhood principle was thus expanded in this case. The concept of negligence was

[27] Caparo Industries Plc. v. Dickman and Others [1990] UKHL 2, [1990] 2 AC 605
[28] Tan Keng Feng, 'The Three-Part Test: Yet Another Test Of Duty in Negligence' 1989) 31 Mal. L. R. 223, 224 <https://law.nus.edu.sg/sjls/wp-content/uploads/sites/14/2024/07/1285-1989-31-mal-dec-223.pdf>
[29] Mihailis E. Diamantis, 'Vicarious Liability for AI' (2023) 99 Ind LJ 317
[30] Shraddha Kamatagi, 'Accountability Beyond Humans: Tortious Liability Frameworks for Artificial Intelligence in India' (2025) 2 (8) IJLRA 4 <https://www.ijlra.com/uploads/1990605102.pdf>
[31] *Home Office v. Dorset Yacht Co Ltd.* [1970] UKHL 2, [1970] AC 1004

expanded further in Hedley Byrne v. Heller, where Lord Morris observed that if someone who possesses a "special skill" applies it for purposes external to the contract to assist another person, and the other person relies upon such skill, a duty of care will arise. Further, Lord Wilberforce devised the "Ann's two-stage test" in 1977.[32] According to this, it is first required to satisfy the "neighbour principle" to establish a prima facie duty of care, and then analyse whether there are any reasons, or policy considerations, that such a duty should not exist. However, this test overlooked the presumption of innocence ("innocent until proven guilty"), as once the plaintiff satisfied the first test based on "foreseeability of harm", the burden to disprove liability shifted to the defendant.[33] In Yuen Kun Yeu v. Attorney-General of Hong Kong,[34] Lord Keith rejected this test by saying that "for the future it should be recognised that the two-stage test in Anns is not to be regarded as in all the circumstances a suitable guide to the existence of a duty of care."

Thus, this test was ultimately overruled, and the current test to establish negligence was formulated on 8 February 1990, when the House of Lords delivered a landmark verdict in Caparo Industries Plc. v. Dickman and Others.[35]

This case revolved around an action brought by a public limited company, the plaintiffs, that took over F. Plc. The action was against the auditors of the company, and the plaintiffs contended that they were negligent in carrying out audits and making reports, as they breached their duty of care to the plaintiffs, who were both shareholders and potential investors, in respect of the certification of accounts. They projected F. Plc.'s profits higher than they actually were, and this projection was relied upon by the plaintiffs. The House of Lords held that the "liability for economic loss due to negligent misstatement" was confined to cases where some advice had been given to a person for a specific purpose known to the advisor and where the person relied upon such advice and acted to his detriment. Here, the auditors' duty of care was only extended to the plaintiffs as shareholders, and not as potential investors and buyers of the F. Plc's shares, but later, even this was rejected.

Foreseeability could be clearly established in this case; hence, it was ruled out for discussion, but the problem rested with ascertaining "proximity". Since this case was of negligent

[32] *Anns v. Merton London Borough Council* [1977] UKHL 4, [1978] AC 728
[33] Moosh, 'Duty of Care' (*Medium,* 24 June 2020) <https://mooshii.medium.com/duty-of-care-ea2dc6478f85 > accessed 10 February 2026
[34] *Yuen Kun Yeu v. Attorney-General of Hong Kong* [1987] Sing JLS 308
[35] *Caparo Industries Plc. v. Dickman and Others* [1990] UKHL 2, [1990] 2 AC 605

misstatement, the following were required to ascertain proximity: a) the maker of the statement intended for the person to act upon it, b) in a specific transaction, and c) for the purpose for which the statement was made. Here, no special relationship existed between Caparo and the defendants, as the opportunity to buy the shares was provided to the world, and the reports were not made by the auditors for a specific purpose. Moreover, the duty of care was limited to "every shareholder", and not an "individual shareholder" like Caparo. The House also pointed out that merely because a strong foreseeability may point towards proximity, they both cannot be equated. It is also pertinent to mention that the adjudicators in this case rejected the prevailing "two-stage" notion given by Lord Wilberforce.

According to Lord Bridge of Harwich,

> "…in addition to the foreseeability of damage, necessary ingredients in any situation giving rise to a duty of care are that there should exist between the party owing the duty and the party to whom it is owed a relationship characterised by the law as one of 'proximity' or 'neighbourhood' and that the situation should be one in which the court considers it fair, just and reasonable that the law should impose a duty of a given scope upon the one party for the benefit of the other."

He, therefore, explained that there were three elements required for a duty of care to exist: 1) reasonable foreseeability, 2) proximity between the parties, and 3) a fair, just, and reasonable duty. It was this threefold test that was utilised in the Caparo case to ascertain whether there was a duty of care owed by the defendant or not. This case presently stands as one of the most studied and referred cases, as it is credited with the creation of this 'three-leg test', also known as the "Caparo test", to establish the duty of care under the tort of negligence adequately.[36] Such a duty of care is the basic prerequisite to constitute the other two essentials of negligence, which are breach of duty, as well as the resulting damage.

Applicability of this Test to AI: Certainly, Winfield's idea of the basic four requirements for negligence stands as supreme, but there is a need for more elaborate criteria or steps to ascertain liability for negligence, owing to the complication of laws and circumstances. This is why the Courts have taken the pains to keep reinterpreting the existing laws and evolving the scope of negligence to suit the societal needs, as discussed previously. To analyse the scope of

[36] Pallavi Agarwal, 'Caparo Test' (2021) 2 Indian JL & Legal Rsch 1

negligence by AI, it is thus important to apply the present, most applicable test, i.e., the Caparo test.

Foreseeability of Damage: The question of foreseeability actually arises at different stages in determining negligence, but its widest scope lies at the first stage, which involves asking the preliminary question, "Was some kind of harm actually foreseeable?" The term "foreseeability" essentially refers to the notion that a person must be able to foresee that their act or omission would have a specific result.[37] Foreseeability is, in fact, one of the greatest challenges that AI poses for tort law.[38] Certainly, AI cannot achieve human-level consciousness,[39] but in the future, there is a chance of Artificial General Intelligence (AGI), also sometimes known as "strong AI," coming into the picture, which would inevitably require an expansion of the range of what is considered foreseeable presently. Till then, however, the foreseeability categories must not be changed.[40] This is to say that the victims can sue against AI defaults under existing laws only, but the problem arises in who can be sued- the AI or the companies?

There can only be two possibilities while determining the applicability of this foreseeability: either AI is considered to be a separate entity, or it is assumed that its developers regulate all results provided by AI. If AI is considered to be a separate entity, foreseeability would have to be determined on the part of the AI system. The duty of care under negligence either requires "ordinary care" or "professional care" on the part of the persons. But if AI is not even considered a separate person and is deemed to have no thinking capability of its own, it becomes difficult to determine whether AI could actually foresee any harm. There certainly are cases where the human agencies play a role in exercising reasonable care, and any breach of such care would be considered wrong,[41] but in recent times, there are also a large number of cases that go beyond the thinking capabilities of humans. For example, in 2016, AlphaGo, an artificial intelligence, beat "one of the best players of perhaps the most complex game ever

[37] Vanshika Agarwal, 'Critical Analysis of the Concept of Foreseeability in Torts of Negligence' (2023) 6 (1) International Journal of Law Management & Humanities 1668 <https://doij.org/10.10000/IJLMH.114167> accessed 11 February 2026
[38] Andrew D. Selbst, 'Negligence and AI's Human Users' (2020) 100 Boston University Law Review 1315, 1342 <https://www.bu.edu/bulawreview/files/2020/09/SELBST.pdf> accessed 12 February 2026
[39] 'Artificial Intelligence Is Nothing Like a Brain — or a Mind' (*Mind Matters*, 7 May 2025) <https://mindmatters.ai/2025/05/artificial-intelligence-is-nothing-like-a-brain-or-a-mind/> accessed 12 February 2026
[40] Andrew D. Selbst, 'Negligence and AI's Human Users' (2020) 100 Boston University Law Review 1315, 1344 <https://www.bu.edu/bulawreview/files/2020/09/SELBST.pdf> accessed 12 February 2026
[41] Ibid at 1345

devised by humans."[42] The AI acted as if it actually understood the game of Go, and it acted in ways no person could imagine. Another example of an AI made by the same company, i.e., DeepMind, is AlphaFold, which solved one of the greatest problems in biology by predicting the 3-D structures of "almost every known protein".[43] These developments are certainly positive, but there are recorded cases of AI negligence as well, and they pose actual concerns for humans. For example, in 2025, an AI coding system built by Replit was accused of generating around 4000 fake users and deleting a live database. According to Jason Lemkin, this AI system modified code without permission and ignored repeated instructions.[44] This brings the second possibility into the picture: considering that all results of AI are regulated by its developers. Various cases, like the AlphaGo case, reveal that AI actually works in ways that go beyond human capabilities. In such situations, if AI does something or provides a result that harms the user of AI, and there was nothing wrong on their part, the question of who is to be made liable arises. In the medical field, for example, the use of Large Language Models by medical professionals will most likely be seen as a "third-party medical guidance," and the standard of care will be according to the prevailing standards only. Here, it is the physicians who are made responsible for AI use, and the third party "may" be made responsible. Thus, neither the AI models nor the developers can be made responsible, and the ultimate burden rests on the physician.[45] The black box problem of AI poses a major problem here. AI systems train themselves and might reach a point of 'Technological Singularity' (by 2045, as predicted by Ray Kurzweil), brought about by AI systems surpassing human cognitive capabilities.[46] It thus becomes extremely difficult to prove foreseeability of damage as the workings of AI systems cannot be completely decoded, and the actions of AI may go beyond human agency. Thus, though there is no requirement to change any categories or range of foreseeability,

[42] Cade Metz, 'What the AI Behind AlphaGo Can Teach Us About Being Human' (*Wired,* 19 May 2016) <https://www.wired.com/2016/05/google-alpha-go-ai/> accessed 12 February 2026

[43] Tanya Lewis, 'One of the Biggest Problems in Biology Has Finally Been Solved' (*Scientific American,* 31 October 2022) <https://www.scientificamerican.com/article/one-of-the-biggest-problems-in-biology-has-finally-been-solved/> -accessed 12 February 2026

[44] 'AI goes rogue: Replit coding tool deletes entire company database, creates fake data for 4,000 users' *The Economic Times* (22 July 2025) <https://economictimes.indiatimes.com/news/new-updates/ai-goes-rogue-replit-coding-tool-deletes-entire-company-database-creates-fake-data-for-4000-users/articleshow/122830424.cms?from=mdr> accessed 12 February 2026

[45] David O. Shumway, Hayes J. Hatman, 'Medical malpractice liability in large language model artificial intelligence: legal review and policy recommendations' (2024) 124 (7) J Osteopath Med. 287, 288 < https://doi.org/10.1515/jom-2023-0229 > accessed 12 February 2026

[46] Tim Mucci, 'What is the technological singularity' (*IBM Think)* <https://www.ibm.com/think/topics/technological-singularity> accessed 12 February 2026

ascertaining proper foreseeability is a difficult and almost impossible task within the present applicability of this test.

Proximity or the 'Neighbourhood': Proximity, in simple terms, refers to the "nearest cause which is responsible for injury", and any harm caused outside the scope of such proximity will not be actionable.[47] The concept of foreseeability and proximity was formulated by Lord Atkin as the "neighbour principle" in Donoghue v. Stevenson, discussed before. Such proximity may be either physical, causal, assumed, or circumstantial. In Hedley Byrne v. Heller, the concept of "special relationship" was devised, where negligence involved a "negligent misrepresentation" to be made, along with certain proximity. This proximity included the following: i) the party relied upon the representation made, ii) the party making the representation had an inferred assumption of responsibility, and iii) there was foreseeability of financial damage. On analysing this interpretation of proximity, it becomes apparent that foreseeability must necessarily be determined before contemplating proximity, as sometimes the mere proof of foreseeability sufficiently proves proximity as well. In Jaensch v. Coffey,[48] Deane J provided different meanings of the term 'proximity', one of which involved "designating a general limitation on reasonable foreseeability" through a relationship that must exist between the plaintiff and defendant before any duty arises. Thus, it is also required to establish a relationship, either "special" or like that of a "neighbour", to establish proximity.

AI systems cannot reasonably be assumed to have an understanding of the reasonable foreseeability of potential harms caused to a neighbour. Normally, developers and deployers of AI owe a duty of care to its users, especially when it is deployed in "high-risk domains" like healthcare and transportation.[49] The liability thus shifts away from AI systems. The black box problem again poses an issue as AI developers may possess no means to ascertain or foresee the results provided by AI systems, especially in cases where these deep-learning systems train themselves, without human intervention. Furthermore, even if AI is considered to possess special skills, in cases where negligence is alleged, the systems cannot be deemed to have any assumed responsibility or foreseeability of damage. In fact, it is the user who will be considered

[47] R. Vandhana Prabhu and Priya Darshani, 'A TEST OF PROXIMITY AND FORESEEABILITY WITH RESPECT TO THE TORT OF NEGLIGENCE: AN INTERNATIONAL PERSPECTIVE' (2018) 120 (5) International Journal of Pure and Applied Mathematics 453, 456 <https://acadpubl.eu/hub/2018-120-5/5/430.pdf> accessed 13 February 2026

[48] *Jaensch v. Coffey* (1984) 155 CLR 549

[49] Shraddha Kamatagi, 'ACCOUNTABILITY BEYOND HUMANS: TORTIOUS LIABILITY FRAMEWORKS FOR ARTIFICIAL INTELLIGENCE IN INDIA' (2025) 2 (8) IJLRA 4,5 <https://www.ijlra.com/uploads/1990605102.pdf> accessed 13 February 2026

at fault for blindly relying upon the skills of AI, even though various cases (the AlphaGo case, for example) actually reveal how AI possesses human-like, or even beyond human-like, capabilities.

The problem of ascertaining proximity for AI systems is rooted in the first stage of "foreseeability". It primarily hinges on the amount of control its developer has on the AI system-if it is more, the developer will have proximity to the user, and can be said to owe a duty of care, but if it is less, or even none, liability rests with practically no one as per the present laws, and dodges between the users, the developers, and the companies at the discretion of the courts.

Fair, Just, and Reasonable Duty: The third test of a duty being 'fair, just and reasonable' is more like a 'policy prong', similar to the second part of Wilberforce's Anns test. This part focuses on policy considerations, i.e., seeing whether the duty imposed by law "ought to be negatived or ousted by policy considerations."[50] The House of Lords in Mitchell v. Glasgow City Council[51] provided clarifications to the Caparo test and held that mere establishment of foreseeability does not create a duty of care, and it is important to see that the duty is 'fair, just and reasonable'. This test, thus, certainly performs a necessary role in the law of negligence. However, this test of reasonability is not always applied uniformly in all cases, and it primarily lies at the discretion of the courts to determine whether a duty of care exists or not.[52] Thus, determining the existence of such a duty is extremely subjective when it has to be determined with respect to this test. As critics of the Anns test also point out, a liberal application of this test could lead to a "floodgate" of claims of negligence. With the first two tests of "foreseeability" and "proximity", this paper addressed the problems of ascertaining AI liability. However, with this test, suppose the first two tests are satisfied, the courts are likely to hold companies and developers of AI liable even if a proper standard of care had been taken by them. This concern primarily stems from the unwillingness of courts to address the black box problem due to possible intellectual property violations, and also because the level of

[50] Joost Blom, 'Do We Really Need the Anns Test for Duty of Care in Negligence?' (2016) 53 Alta L Rev 895, 902
[51] *Mitchell and Another v Glasgow City Council* [2009] UKHL 11, [2009] 1 AC 874
[52] Pallavi Agarwal, 'Caparo Test' (2021) 2 Indian JL & Legal Rsch 1, 2

complexity with which deep-learning systems operate makes it difficult to decipher their logic behind making decisions.[53]

Thus, the Caparo test has been getting increasingly fragile and vulnerable before AI's black box. As these deep-learning systems sometimes tend to surpass human cognitive abilities, proving foreseeability in case of AI becomes difficult, as developers are unable to predict the outputs they did not program. This further creates a proximity vacuum. Since the developers lack control over the outputs, the "neighbourhood" connection breaks, and establishing proximity gets problematic. Finally, the test of fairness, justice and reasonableness almost becomes a policy gamble, as courts may either unfairly penalise developers, or leave victims without any judicial recourse, as AI's decision-making remains a mystery due to its black box.

SUGGESTIONS TO RESOLVE THIS ISSUE

Based on its analysis, this article proposes three solutions to the primary issue.

Checks on AI Systems: A key aspect in determining the validity of AI is that its users are able to predict its actions ex-ante, and thus, more predictable AI systems are usually preferable in all sectors. Moreover, AI predictability can make it easier to determine foreseeability by any diligent method, and not necessarily by its designers or systems.[54] Ensuring human control over AI systems also prevents mistakes and increases the accuracy and efficiency of the "human-AI teams".[55] Thus, creating more predictable AI systems might be a way to make foreseeability easier. This differs from explainable AI (xAI), which is yet another solution for the black box problem while determining foreseeability. Predictable AI seeks to anticipate the observable indicators, while xAI aims to find out how an AI system arrives at a particular decision.[56] However, there is still a requirement for more technological developments to enforce this, as even though xAI models make it easier to trust AI outcomes, it is not easy to turn all AI systems into a white box. It is easier in traditional AI models by sharing source codes, but not in deep-

[53] Amandeep Singh and Janees Rafiq, 'IMPLICATIONS OF BLACK BOX DILEMMA IN THE INDIAN LEGAL SYSTEM' (2025) 4 (3) Journal of Legal Research and Juridical Sciences 1264, 1269 <https://jlrjs.com/wp-content/uploads/2025/06/111.-Amandeep-Singh.pdf> accessed 15 February 2026
[54] Lexin Zhou *et al*., 'Predictable artificial intelligence' (2026) 353 Artificial Intelligence 1 <https://www.sciencedirect.com/science/article/pii/S0004370226000172> accessed 15 February 2026
[55] Serhiy Kandul *et al.*, 'Human control redressed: Comparing AI and human predictability in a real-effort task' (2023) 10 Computers in Human Behaviour Reports <https://www.sciencedirect.com/science/article/pii/S2451958823000234> accessed 15 February 2026
[56] Lexin Zhou *et al*., 'Predictable artificial intelligence' (2026) 353 Artificial Intelligence 1 <https://www.sciencedirect.com/science/article/pii/S0004370226000172> accessed 15 February 2026

learning systems.[57] In recent years, researchers at Anthropic have increased AI predictability by around 70% by deploying an autoencoder to overcome polysemanticity. Such predictability can help overcome the first part of the Caparo test. i.e., "foreseeability", as actions of AI systems become more predictable. In February 2026, OpenAI introduced 'Lockdown Mode' in ChatGPT, which restricts how this system interacts with external systems.[58] Similar restrictions may be placed on AI platforms to control the decision-making processes of such deep-learning systems and keep them in check. This step can prove to be useful for courts to easily establish foreseeability in cases of AI systems by incorporating a novel understanding of the black box problem.

Establishing Vicarious Liability for AI Developers/ Companies: The second suggestion is based on constructing a "vicarious liability model" for AI systems. A similar variant of the criminal law doctrine of solicitation liability is part of the civil law in some places. If a programmer designs a program to commit an offence, he has the mens rea required under criminal law. Even though the actual offence may be committed by an AI entity, there is no mental attribute to AI. In such cases, it is the end-user who is deemed the perpetrator, due to the actus reus. Hence, the user of the AI entity might be considered the "perpetrator-via-another".[59] However, this "liability model" would be suitable in case the AI system is either old or very basic. If the users, who may be the victims of negligence under tort law, are deemed responsible for the acts of the AI system, the primary motive of civil law. i.e., restoration, would be lost. In case of negligence, it is assumed that the programmers or users of AI should have known about the probability of the AI committing an offence, even though they did not actually know about it.[60] This "negligence-based model" does not place any limits on the connections that could implicate a defendant, i.e., this can apply to any negligent party equally.[61] This would be detrimental as a black box often prevents predictability and analysis of results. In case of very advanced AI systems, however, the AI entities may even be held

[57] Matthew Kosinski, 'What is black box AI?' (*IBM Think*, 29 October 2024) <https://www.ibm.com/think/topics/black-box-ai> accessed 16 February 2026
[58] Mark Tarre, 'OpenAI unveils Lockdown Mode to counter prompt attacks' *SecurityBrief Australia* (16 February 2026) <https://securitybrief.com.au/story/openai-unveils-lockdown-mode-to-counter-prompt-attacks> accessed 16 February 2026
[59] Gabriel Hallevy, 'The Criminal Liability of Artificial Intelligence Entities - from Science Fiction to Legal Social Control' (2016) 4 (2) Akron Intellectual Property Journal 171, 180 <https://ideaexchange.uakron.edu/cgi/viewcontent.cgi?article=1037&context=akronintellectualproperty> accessed 16 February 2026
[60] Ibid at 183
[61] Mihailis E. Diamantis, 'Vicarious Liability for AI' (2023) 99 Ind LJ 317, 322 <https://heinonline.org/HOL/Page?handle=hein.journals/indana99&id=330> accessed 15 February 2026

criminally liable.[62] The main problem here rests with ascertaining who would actually be made responsible for providing compensation to the victim and bearing the liability. A comprehensive vicarious liability model could be helpful in such cases, a necessary prerequisite to which is considering AI to be considered an "electronic person" only for adjudication purposes. If an AI is considered an agent, and the deployers are principals, the deployers must be made vicariously liable for negligence by the AI's agent. Their liability, however, must be reduced based on AI's black box. If it is established that the system yielded detrimental results due to its own deep-learning mechanisms, then the liability must be reduced automatically. However, if the black box had been inserted wilfully with no rational justification, or if there is "solicitation", the developer of the AI must be made jointly liable, i.e., the deployer must share liability with such a developer. The deployer needs to be made liable as they also owed a duty of care towards the users, and it was they whom the users trusted. This may also be tackled using strict liability mechanisms, as discussed in the next section. The first suggestion to incorporate predictable and explainable AI mechanisms, therefore, also becomes a prerequisite to this suggestion, as these mechanisms are crucial to be able to establish the level of liability adequately.

Incorporating Strict Liability Rules within AI: The last suggestion stems from the European Union's EU Act, 2024, which highlights the advantages that a single liability point (typically the developers) can offer when ascertaining liability in the AI chain.[63] The Act mandates the provision of technical documentation and instructions for use by the General Purpose AI (GPAI) providers, along with a summary of the content used for training of the AI. Furthermore, the Act provides for the establishment of proper risk management systems for high-risk AI systems.[64] There has, thus, been an attempt to establish strict liability for developers and deployers regarding AI systems. Similar to the second suggestion, this would eliminate liability from the user and would not additionally require combating the legal battles of attributing AI the status of a legal entity. Moreover, strict liability rules would help overcome

[62] Gabriel Hallevy, 'The Criminal Liability of Artificial Intelligence Entities - from Science Fiction to Legal Social Control' (2016) 4 (2) Akron Intellectual Property Journal 171, 187 <https://ideaexchange.uakron.edu/cgi/viewcontent.cgi?article=1037&context=akronintellectualproperty> accessed 16 February 2026

[63] Kamya Pandey, 'EU Study Suggests Strict Liability For High-Risk AI: What This Means' (*Medianama*, 4 August 2025) <https://www.medianama.com/2025/08/223-eu-study-strict-liability-high-risk-ai/> accessed 16 February 2026

[64] Regulation 2024/1689/EC of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L2024/1689

the features of "autonomy, imperfection, unpredictability, and opacity" within AI, which pose challenges for establishing negligence. A "rebuttable presumption of negligence" that shifts the burden of proof to the defendant might also help maintain the "innocent until proven guilty" principle, while ascertaining liability aptly. This suggestion does not originate within a vacuum, as existing product liability rules encompass "manufacturing defects" and the associated strict liability rules to address harms that arise from any flaws during the manufacturing process. Thus, any design defects or intentional addition of black boxes, leading to negligence, can be considered equivalent to manufacturing defects, thereby warranting "strict liability" for the developers.[65] Both product liability and negligence can be utilised to provide a robust system to address AI negligence.

Strict liability also circumvents the issues with the applicability of the Caparo test. It is not necessary to establish whether the developer could foresee any harm or not, and they would be held liable regardless of technical limitations. It is, however, not to say that despite the presence of a black box in AI, innocent developers shall be made liable, but that while programming the AI system, it was their responsibility to create an interpretable AI. Developing a black box AI tends to be easier to develop as compared to an interpretable AI, which requires "domain expertise and specialised talent".[66] Thus, developers could be made strictly liable for having failed to curtail the harms caused by AI systems adequately. In such cases, establishing an alternative "standard of care" might be considered in cases where the black box is "organic" and not inserted by the programmers. Similar to the Bolam test,[67] which attributes the standard of a "reasonable body of medical men" to medical practitioners, there could also be a standard of a "reasonable body of developers" to establish where they could be held liable for breaching their duty of care. Laws like the EU AI Liability Directive and the UK's AI Regulation Bill have categorised AI into high-risk and low-risk AI, providing for strict liability and negligence rules for each category, respectively, which implies that a distinction between AI, based on risk categories, can improve the assessment of AI liability.

[65] Maarten Herbosch, 'How Existing Liability Frameworks Can Handle Agentic AI Harms' (*Lawfare*, 3 December 2025) <https://www.lawfaremedia.org/article/how-existing-liability-frameworks-can-handle-agentic-ai-harms> accessed 17 February 2026

[66] Mihailis E. Diamantis, 'Vicarious Liability for AI' (2023) 99 Ind LJ 317, 318 <https://heinonline.org/HOL/Page?handle=hein.journals/indana99&id=330> accessed 17 February 2026

[67] *Bolam v Friern Hospital Management Committee* [1957] 1 WLR 582

CONCLUSION

The evolution of AI from being a mere trend-finding machine to becoming a complex thinking machine has led to a tendency for the traditional foundations of tort law to be disrupted. This article has demonstrated that the Caparo test, which has been the judicial standard for establishing a duty of care for over thirty years, is becoming increasingly vulnerable before AI's black box. The polysemanticity and the unpredictable and autonomous nature of deep-learning systems create opacity that undermines the three-fold requirements of "foreseeability", "proximity", and a "just, fair and reasonable duty". The first prong falters as AI systems presently tend to surpass human cognitive abilities, making it nearly impossible for developers to predict specific detrimental results. This lack of predictability subsequently creates a proximity vacuum, as the long-standing "neighbourhood" connection between the manufacturer and the end-user is severed by the AI's independent decision-making processes. Lastly, the third test risks becoming a policy gamble, with a potential of leaving victims without recourse or of unfairly penalising developers for outputs provided by independent, or "organic" black boxes. Therefore, a multifaceted approach is essential to tackle the gap between technological and legal advancements. The integration of AI systems, opposite to black boxes, like Predictable AI and xAI, becomes necessary to reinstate foreseeability. Furthermore, the law must attempt to move towards a suitable "vicarious liability model" that treats AI as a legal agent, or adopt strict liability rules for high-risk AI systems. Thus, by attempting to shift the focus from intent and foresight to risk management, and by recognising the presence of issues like the black box, the legal system can ensure that the law of torts continues to provide justice even in an automated age. In conclusion, the law of torts must continue to evolve as a living law to ensure that the black box does not become a shield against accountability, while at the same time it does not prove to be detrimental for genuine victims and innocent persons.